16. Speech Communication

Academic and Research Staff

Prof. K.N. Stevens, Prof. J. Allen, Prof. M. Halle, Prof. S.J. Keyser, Prof. V.W. Zue, Dr. R. Carlson[11], Dr. T. Carrell, Dr. C. Chapin Ringo, M. Cohen, Dr. C. Connine, Dr. B. Delgutte[12], Dr. M. Danly[13], Dr. R. Goldhor[14], Dr. B. Greene[12], Dr. F. Grosjean[15], Dr. S. Hawkins[16], Dr. R.E. Hillman[17], E.B. Holmberg[18], Dr. A.W.F. Huggins[19], S. Hunnicutt[12], Dr. H. Kawasaki[20] Dr. D.H. Klatt, K. Krohn, Dr. L.S. Larkey[14], Dr. B. Lherm[12], Dr. J. Locke[21] Dr. S. Maeda[12], Dr. J.I. Makhoul[12], Dr. L. Menn[22], Dr. P. Menyuk[23], Dr. J.L. Miller[15], Dr. J.S. Perkell, Dr. D. Pisoni[12], Dr. P.J. Price[19], Dr. S. Seneff, Dr. S. Shattuck-Hufnagel, R. Schulman[12], Dr. J. Vaissiere[12], Dr. K. Yoshida[12]

Graduate Students

A. Alwan, M. Anderson, C. Aoki, C. Bickley, T. Cass, F. Chen, S. Cyphers, N. Daly, S. Dubois, C. Espy-Wilson, J. Glass, C. Huang, D. Huttenlocher, L. Lamel, N. Lauritzen, H. Leung, K. Moore, L. Pfeifer, J. Pitrelli, M. Randolph, C. Shadle, T. Wilson

---

[11] Guest

[12] Visiting Scientist

[13] Director of Speech Product Development, Wang Laboratories

[14] Staff Member, Kurzweil Applied Intelligence

[15] Associate Professor, Department of Psychology, Northeastern University

[16] Postdoctoral Fellow, Haskins Laboratories

[17] Assistant Professor, Department of Speech Disorders, Boston University

[18] Research Scientist, Department of Speech Disorders, Boston University

[19] Staff Member, Bolt, Beranek and Newman, Inc.

[20] Staff Member, Voice Processing Corporation

[21] Director, Neurolinguistics Laboratory, Massachusetts General Hospital, Institute of Health Professions

[22] Associate Professor, Department of Neurology, Boston University School of Medicine

[23] Professor of Special Education, Boston University

C.J. LeBel Fellowship
Kurzweil Applied Intelligence
National Institutes of Health (Grants 5 T32 NS07040 and 5 RO1 NS04332)
National Science Foundation (Grant BNS84-18733)
Systems Development Foundation
U.S. Navy - Office of Naval Research (Contract N00014-82-K-0727)
Kenneth N. Stevens, Dennis H. Klatt, Joseph S. Perkell, Stefanie Shattuck-Hufnagel, Stephanie Seneff, Victor W. Zue

16.1 Speech Recognition

The overall objectives of our research in machine recognition of speech are:

a) to carry out research aimed at collecting, quantifying, and organizing acoustic-phonetic knowledge, and

b) to develop techniques for incorporating such knowledge, as well as other relevant linguistic knowledge, into speech recognition systems.

During the past year, progress has been made on several projects related to these broad objectives.

16.1.1 Applying Auditory Models to Speech Recognition

The human auditory system is an existing speech recognizer with excellent performance. If we could build computer models that adequately reflect the transformations that take place in the ear, utilizing research results from auditory physiologists and psychophysicists, then the resulting "spectral" representations may be superior to other representations for computer speech recognition. Whether such models will ultimately yield better computer speech recognition is not yet known. However, we believe that a worthwhile approach is to build speech recognition systems that incorporate an auditory-based front-end analysis scheme. Such a signal representation may make it easier to identify those aspects of the signal that are relevant for recognition, and may lead to valid ideas for appropriate strategies for later stage processing.

In the last two years, we have developed one such auditory-based model. The analysis system consists of a set of 40 independent channels, which collectively cover the frequency range from 130 to 6400 Hz. Each channel consists of a linear critical-band filter, followed by a nonlinear stage that models the hair-cell/synapse transformation. The outputs of this stage are delivered to two parallel noninteracting modules, one of which determines the envelope response, corresponding to "mean rate response" of auditory neurons, and the other of which measures the extent of dominance of information at the characteristic frequency of the linear filter; i.e., determines the "synchronous response."

We believe that each of these representations is useful for a different aspect of the speech recognition problem. The hair-cell envelope response tends to show sharper onsets and offsets than those produced after only the linear stage. It therefore should be useful for determining acoustic boundaries. Furthermore, due in part to saturation phenomena, steady-state prominences tend to become broader in frequency, and this may make it easier to group segments into broad acoustic classes. We have, in fact, had encouraging success in utilizing the mean-rate response for detecting acoustic boundaries between vowels and adjacent nasal consonants.

The other representation resulting from our auditory model, the output of the synchrony module, is expected to be useful for making fine phonetic distinctions among word candidates within the same acoustic class. The synchrony module measures the extent of dominance of information near the filter center frequency in the channel response. The outputs of this stage generally show narrow peaks at the formant frequencies, and thus should be suitable for distinguishing between competitors within the same broad category. To test this assumption, we have begun to apply the outputs of the synchronous response stage to the task of vowel recognition.

16.1.2 Formalizing the Process of Spectrogram Reading

Over the past four decades the spectrogram has been the single most widely used form of display in speech research. Recently, the spectrographic display took on added significance when experiments demonstrated that the underlying phonetic representation of an unknown utterance can be extracted almost entirely from a visual examination of the speech spectrogram. These experiments stirred renewed interest in acoustic-phonetic approaches to speech recognition, and supported the speculation that better front-ends may be constructed if we can learn the phonetic decoding procedure used by human experts.

Over the past year we have initiated two projects aimed at formalizing the process of spectrogram reading. The first is directed toward signal processing and feature extraction, and the second focuses on knowledge representation and integration. Protocol analysis of spectrogram reading reveals that the decoding process calls for the recognition and integration of a myriad of acoustic patterns. In order to develop a system that utilizes such knowledge, one must first be able to extract these acoustic patterns. The aim of our first project is to capture the essential acoustic patterns of a spectrogram so that these abstracted patterns may be used to characterize and recognize different speech sounds. Traditional descriptions of acoustic-phonetic events based on formant frequencies are often inadequate because the formants cannot always be resolved reliably. Thus visual characterizations may provide an alternative, and perhaps more effective, description.

Our approach to visual pattern extraction borrows liberally from techniques used in vision

research. The spectrographic image is processed through a set of edge detectors, whose outputs are then combined by applying scale-space filtering techniques, along with constraints imposed by our knowledge of the speech production mechanism. This procedure results in a representation of the spectrograms as a set of two-dimensional objects that characterize the formant patterns and general spectral properties for vowels and consonants. As a validation of the approach, a limited vowel recognition experiment was performed on the "object" spectrograms. Preliminary results, based on the recognition of vowels from the visual objects, show that this processing technique retains relevant acoustic information necessary to identify the underlying phonetic representation.

Spectrogram reading depends on the recognition and integration of many acoustic cues. Some of these cues are relatively easy to identify, while others are not meaningful until the relevant context has been established. One must selectively attend to many different acoustic cues, interpret their significance in light of other evidence, and make inferences based on information from multiple sources. The discovery of the acoustic cues and, equally importantly, the determination of control strategies for utilizing these cues, are the keys to high-performance phonetic recognition. The aim of our second project is to investigate the procedures used by spectrogram readers to analyze a set of acoustic cues and arrive at a phonetic judgment. Our experience with spectrogram reading suggests that the reasoning process can be naturally expressed as a series of production (or *if-then*) rules, where the preconditions and conclusions may be phonetic features or acoustic events. Since the acoustic-phonetic encoding is highly context-dependent and redundant, we must be able to entertain multiple hypotheses and check for consistency. Acoustic features are often expressed in a qualitative manner and described as being present/absent, and having values such as high/mid/low, or weak/strong. Thus in order for the computer to mimic the performance of spectrogram readers, it must be able to deal with qualitative measures in a meaningful way.

We have attempted to incorporate our knowledge about the spectrogram-reading process into a knowledge-based system that mimics the process of feature identification and logical deduction used by experts. The knowledge base explicitly represents the expert's knowledge in a way that is easy to understand, modify, and update. Our first attempt in this direction makes use of an existing *Mycin*-based expert system to investigate stop consonants in singleton and in clusters. Preliminary recognition results, based on singleton stops from multiple talkers, indicate that system performance is comparable to that of human experts.

16.1.3 Development of Tools for Research in Speech Recognition

There are a number of tasks that speech researchers commonly perform: recording and digitizing utterances, defining and computing various attributes of the speech signal, displaying and performing interactive measurements of these attributes, using a large speech database to statistically analyze the interrelation between acoustic and phonetic events, studying the

phonological properties of language at the symbolic level by using large lexicons and/or printed text, and interactively synthesizing speech so as to determine the relative merits of acoustic cues for phonetic contrasts. In order to perform such tasks more easily, we have designed a speech research workstation based on a Symbolics Lisp Machine, for which we are developing a special set of research tools.

Spire (Speech and Phonetics Interactive Research Environment) is a software package that allows users to digitize, transcribe, process, and examine speech signals. Speech can be sampled at any rate up to 75 kHz, and a variety of computations can be performed on the signal with a minimum amount of user programming. A variety of attributes can then be displayed on the Lisp Machine monitor, including the wide-band spectrogram of the utterance, the zero-crossing rate, the original waveform, the narrow-band spectrum, and the LPC spectrum. *Spire* also provides a convenient mechanism for entering an orthographic or phonetic transcription and time-aligning it with the speech waveform. Users have the flexibility to manipulate these displays as well as to generate hard-copy printouts of what is on the screen.

Search (Structured Environment for Assimilating the Regularities in speeCH) is an interactive tool designed for exploratory analysis of speech data. *Search* consists of a set of statistical tools with extensive graphics capabilities for displaying data in various forms, including histograms, scatter-plots, and a bar-like display that shows univariate distributions of data as a function of categorical variables (e.g., speaker sex or phonetic environment). *Search* also features a set of extensible data structures that form a convenient workbench for the design, implementation, and testing of various classification algorithms.

Another software tool we are developing, *ALexiS*, is intended as an aid for displaying and analyzing the constraints of a lexicon. A language is limited not only by the inventory of basic sound units, but also by the frequency of usage and the allowable combinations of these sounds. *ALexiS* enables users to determine the frequency with which sound patterns occur, to study the phonotactic constraints imposed by the language, and to test the effectiveness of phonetic and phonological rules. In addition, users can define new operations and integrate them into the program.

Synth is a speech synthesis facility for studying the relevant perceptual cues in speech. *Synth* implements the Klatt synthesizer with a number of *Spire* computations in the FPS array processor. The Klatt synthesizer allows the user to manipulate parameters such as fundamental frequency; formant frequencies, bandwidths, and amplitudes; amplitudes of voicing, frication, and aspiration; and frequencies and bandwidths of nasal and glottal poles and zeros. The Lisp Machine implementation provides graceful interaction with *Spire*, so, for example, parameters can be traced from natural recordings, and the resulting synthetic stimuli can be readily examined visually and aurally.

16.2 Physiology of Speech Production

16.2.1 Studies of Coarticulation

Situations frequently arise in which the feature specification for a segment is neutral with regard to a particular feature. An example is the feature of rounding in English (and in many other languages), which is specified for certain back vowels, but which appears to be not specified for consonants such as /s t d n k g/. One might expect, then, that in producing an utterance containing a rounded vowel preceded by one or more of these consonants, a speaker is free to anticipate the rounding during the consonant(s). If the consonant occurs between two rounded vowels, one might expect rounding to continue through the consonant from one vowel to the next. These questions were examined in a two-part experiment in which protrusion movements of the lower lip were measured for the vowel /u/. Two speakers of French, a speaker of Spanish, and two speakers of English produced similar sets of nonsense utterances. A bite block was used to eliminate mandible movements as a source of lower lip movements. A detailed examination was made of relationships among events in multiple repetitions of test utterances of the form [uC_nu] and [$V_{unrounded}C_n$u], where $V_{unrounded}$ = [i,*a*] and C_n is one of [t], [nt], and [nst]. Results show that there was relaxation of protrusion during the consonants between the rounded vowels, especially for [s] and [t], manifested by a "trough" in the signal displaying protrusion versus time. For the utterances with an initial unrounded vowel, there was evidence in the protrusion versus time signal of specific, context-dependent protrusion "targets" for the consonants. The protrusion gestures were "partitioned" into components which had approximate synchrony with the acoustic manifestations of the consonants. The data are too limited in scope to suggest a comprehensive explanation of the results. However, they do indicate that there are overlapping effects of adjacent phonetic segments on certain portions of articulatory trajectories ("co-production") and that sounds which are unspecified phonologically for some features may assume positive, context-dependent values of the articulatory correlates of those features. The most important conclusion from this experiment is that further token-by-token examination of movement trajectories in relation to acoustic events is necessary for an in-depth understanding of the relationship between articulatory movements and the underlying phonological structure of utterances.

To understand further the findings from the above study on anticipatory coarticulation, we have continued these experiments with an extensive study of upper lip protrusion movements for the vowel /u/ on one additional subject, a speaker of American English. Test utterances consisted of pairs of words such as "leak hoot" and "leaked coot," each pair embedded in a carrier phrase. Word pairs were chosen so that different combinations of consonants intervene between the two vowels. Protrusion movements and the acoustic signal were recorded, and plots were generated in which movement events for multiple individual tokens can be examined in relation to interactively-determined times of acoustic events (sound segment boundaries). Preliminary qualitative analysis of upper lip protrusion gestures produced results which are in accordance

with those from the study described above. The gestures were partitioned into two components: an initial slow one, with onset of positive velocity occurring during the preceding unrounded vowel, and a later faster one, with its onset marked by the largest acceleration peak, occurring after the end of the unrounded vowel. The temporal and spatial characteristics of the gesture components vary according to the particular sequence of intervocalic consonants, and depend on interactions between the consonants and the surrounding vowels. The results from these two studies support a "hybrid model of anticipatory coarticulation" in which gesture onset times and spatial characteristics are context-dependent, and there is "co-production," i.e., overlapping and summation of multiple influences on articulatory trajectories.

16.2.2 Production of Fricative Consonants

The articulatory basis for the distinction between [s] and [s̬] has been examined in a study in which articulatory data from a dynamic palatograph were collected during the production of coronal fricative consonants. Palatographic and acoustic signals were recorded during continuous fricative sounds produced while the constriction position was gradually changed from a dental to a palatal location. A newly developed computer-based analysis system was used to sample, display, and extract data from the palatographic signal and compare these data with time-synchronized acoustic spectra. The principal peak in the spectrum should be at the frequency of the front-cavity resonance, which decreases during the experimental maneuver. Data analysis and interpretation for two subjects indicate that at some point during this maneuver there is an abrupt drop in the principal spectral peak from the region of F6 or F5 to F3. This rapid change in frequency can be explained by an abrupt increase in front-cavity length that occurs as the sublingual surface breaks contact with the alveolar region on the floor of the mandible. This experiment provides support for the existence of a well-defined boundary between the anterior and nonanterior fricatives, based on articulatory-acoustic relations.

Another aspect of the [s-s̬] articulatory distinction that has been reported informally in the past is that there is a tendency for the lips to be more rounded for [s̬]. Rounding would have the effect of lengthening the front cavity, and would further guarantee a front-cavity resonant frequency that was low enough to fall within the region of the third formant. Experimental data have been collected to determine the validity of this informal observation. Maximum lip protrusion was measured during [s] and [s̬] produced in a variety of environments, including real word combinations and nonsense utterances. Lip protrusion was greater for [s̬] than for [s], by about 0.9 mm on the average. Further analysis of the data examined whether lip protrusion is different when [s̬] is followed by a velar consonant (for which the backed tongue-body position might favor production of a more retracted tongue blade position for [s] and hence less need to protrude the lips) as opposed to a dental consonant [θ] (for which the fronted tongue-blade position might lead to a more fronted tongue-blade position for [s] and hence more need for a protrusion gesture to lengthen the front cavity). Lip protrusion was reliably greater for [s̬] following a fronted consonant [θ] in only one case for one subject, that of nonsense utterances spoken by the

experimenter. The lack of reliable differences for all other cases was interpreted as evidence against the coarticulation hypothesis.

16.3 Speech Planning

Our aim in this series of studies has been to identify the planning units that are represented during the process of phonological planning for speech production, by examining data from phonological or "sublexical" errors that break up words and morphemes. We have been concerned with three aspects of planning representations: first, the planning units themselves; second, the larger structural framework that guides the processing of these phonological elements; and third, the planning process that integrates the units with the framework.

Our studies during the period of this progress report address the question of what successive planning representations look like, and how the phonological planning process makes use of them. In particular, we have looked at the following questions: (1) What are the sublexical error units and what do they reveal about the nature of the planning elements? (2) What are the contextual similarities between pairs of interacting segments, and what do these similarities suggest about the nature of the planning framework? (3) What is the nature of the processing mechanisms that make use of these representations? What are the processing steps involved, in what order do they apply, and what kind of information is available to the processor at each step?

Recent results can be briefly summarized as follows. (1) Planning elements: Polysegmental sublexical error units favor syllabic constituents such as the onset, complex nucleus, and rhyme, suggesting that syllable structure plays a role in planning representations. Vowel errors have been compared with constant errors and found to be similarly constrained by shared distinctive features. (2) Planning framework: The similar position of pairs of error segments in their respective words suggests that the word or morpheme unit is maintained as a structural unit at the point where single-segment errors occur. Lexical stress similarity also plays a measurable but less significant role in consonant errors. For vowel errors, however, primary lexical stress appears to be the most important determinant of confusability in errors. (3) Order of processing mechanisms: Acoustic measurements of the duration of vocalic nuclei after an error which changes the voicing value of a final consonant (e.g., hit had → hid hat) show that these durations are computed *after* single-segment errors occur.

16.4 Studies of the Acoustics and Perception of Speech Sounds

16.4.1 Vowel Perception

We are continuing a series of experiments on the perception of vowels, with the aims of (1) gaining further insight into how the auditory system processes vowel-like sounds, and (2) providing an auditory basis for the classification of vowels in terms of features.

In one set of experiments, vowels with various trajectories for the formants were synthesized in a CVC context, and the effect of trajectory shape on vowel identification was determined. Results showed that perception of vowel height was correlated with a weighted average of the first-formant trajectory through the vowel, with greater weighting applied to the initial part of the vowel. On the other hand, perception of the front-back dimension with a symmetrical consonant context indicated a perceptual overshoot in the second-formant frequency F2, i.e., the backness of the vowel was associated with that of a steady-state vowel for which F2 was greater than the maximum F2 for a concave downward trajectory, and vice versa. These results need to be extended over a wider range of conditions, but they provide tentative evidence for the operation of a different perceptual process for F1 and F2 for vowels.

In another experiment, listeners matched the "openness" of single-formant vowels against five-formant back vowels with various frequencies and spacings of F1 and F2. The fundamental frequency was appropriate for a male voice. The results did not support directly a critical-distance hypothesis in which the matching was between F1 and F2 when the spacing (in Bark) was less than a certain critical value. For high vowels (F1 $<$ 400 Hz, approximately), matching was always to F1, independent of the spacing. For low vowels (F1 $>$ 600 Hz), on the other hand, the matching frequency was always between F1 and F2, independent of the spacing. For the intermediate vowels there were differences between subjects (among the four tested).

16.4.2 Sound Generation in Stop Consonants

It is well known that the sound source at the release of a stop consonant has two components: one is caused by transient excitation of the vocal tract (i.e., a coherent source) and the other by a turbulence noise source. The detailed nature of these sources has not received sufficient attention in the past, and, in particular, there has been little careful study of the coherent component of the source. We have carried out a theoretical study of the temporal and spectral characteristics of these sources for a limited set of rather idealized conditions. Calculations were made for several values of rate of release, glottal opening, and constriction shape. The results of the calculations show that, at the release of a stop consonant like a labial or alveolar, there is an initial transient coherent source, and this is followed by a turbulence noise source that reaches a peak amplitude in the time interval 0.5 to 2 msec. The amplitude of the coherent source can be significant in relation to the noise burst. The coherent source itself has two components: one

due to the negative pressure caused by the rapid release of the articulators, and the other due to the rapid outward flow of air resulting from the intraoral pressure. The first of these components can be present at the release of nasal consonants.

These preliminary model studies need to be extended, and direct comparisons with actual stop releases should be made. The work emphasizes the importance of events during the first few milliseconds following the consonantal release in determining manner and place of articulation. It is leading us to a reexamination of the way we carry out analysis of stop (and nasal) consonants and the way we synthesize these consonants.

16.4.3 Glottal Vibration Source

We have carried out an experimental study in which we measured effects of vocal-tract constrictions on the glottal source from acoustic data and from data from an electroglottograph (EGG) which measures fluctuations in resistance across the neck as the glottis opens and closes. The data show a longer glottal open time for the more constricted configurations, as well as a decreased area of the glottal pulse, less abrupt closing slope of the pulse, and wider bandwidth of the first formant. There was variability in all of the measures across subjects.

In order to study the mechanism by which this interaction occurs, we examined the behavior of a two-mass model of the vocal folds when it is coupled to vocal-tract configurations with various degrees of constriction. The physical parameters of the model were based on data reported in the literature. Attention was focused on obtaining a valid model for the open phase of vibration. There was general agreement between the glottal pulse waveforms generated by the model and those inferred from the speech experiments. In future work we plan to develop this model further, with particular emphasis on the behavior of the vocal folds during the closed phase.

16.4.4 Liquids and Glides: the Feature [-Syllabic]

In connection with our studies of the mechanism of larynx vibration, we have examined the acoustic consequences of producing a consonant-like constriction in the vocal tract while maintaining vocal-fold vibration. In particular, we carried out an acoustic analysis of the sound of the constricted and unconstricted regions of the utterances [ɪwɪ], [ɪjɪ], [ɪrɪ], and [ɪlɪ] (i.e., with liquids or glides forming the constrictions in intervocalic position), produced by six speakers. We used an analysis procedure that allowed us to estimate separately the approximate effects of the constriction on (1) the glottal waveform, (2) the frequency of the first formant, and (3) the bandwidth of the first formant. On the average, all of these consonants produced a decrease in the amplitude of the glottal pulses, a decrease in F1, and an increase in F1 bandwidth. Each of these effects decreases the amplitude of the F1 peak in the spectrum, and thus produces the reduction in intensity that is to be expected for a nonsyllabic segment. The most consistent effect across all six subjects and all consonants is the increase in F1 bandwidth. This increase in bandwidth can be predicted from theoretical considerations (involving the kinetic resistance of

the vocal-tract constriction) if certain assumptions are made about the cross-sectional area of the constriction. A similar bandwidth increase has also been observed for voiced uvular consonants in Arabic, for which there is usually not a decrease in F1 for the consonant in relation to the following vowel. We conclude that the principal mechanism that a speaker uses to obtain an intensity decrease for a liquid and glide (and probably also for other voiced nonnasal consonants) adjacent to a vowel is an increased F1 bandwidth resulting from increased acoustic losses at the supraglottal and glottal constrictions.